A COMPARISON OF DIFFERENT SPECTRAL ANALYSIS MODELS FOR SPEECH RECOGNITION USING NEURAL NETWORKS

Ricardo S. Zebulum †
Marley Vellasco† §

Guy Perelmuter †
Marco Aurélio Pacheco † §

† ICA: Núcleo de Pesquisa em
Inteligência Computacional Aplicada
Departamento de Engenharia Elétrica, PUC-Rio
Rua Marquês de S. Vicente, 225
CEP 22453-900, Rio de Janeiro - Brasil
E-mail: ICA@ele.puc-rio.br

§ Departamento de Sistemas de Computação
Universidade do Estado do Rio de Janeiro
Rua S. Francisco Xavier, 524/5° andar
CEP 20550-013, Rio de Janeiro - Brasil

Abstract

This work presents an application of Artificial Neural Networks (ANN) in Speech Recognition. The aim of this article is to compare the performance of a neural network-based recognition system when using different spectral analysis models. Different sets of coefficients, such as AutoCorrelation and MelCepstron, are extracted from the speech utterances. We have made experiments using, separately, different sets of coefficients as the Neural Network inputs. The development of a hybrid system, combining two different sets of coefficients, has also been performed. The results indicate that the hybrid approach outperforms the other models.

1. INTRODUCTION

This work presents the application of *Artificial Neural Networks* [1, 9] in the problem of speech recognition. The main objective is to compare the performance of a Neural Network-based recognition system when using different techniques to build the spectral vectors. Three different sets of coefficients, the *MelCepstron, Auto Correlation and Reflection* coefficients [2] are used, separately, as the Neural Network inputs. In addition, a combined approach, in which the MelCepstron and the Auto Correlation coefficients are used together as the Neural Network's inputs, is also implemented. The idea behind these experiments is to identify the most suitable spectral analysis model for a Neural Network-based system designed to identify isolated words.

Speech recognition comprises a large number of complex applications [2]. This paper focus on the particular problem of recognizing simple isolated words [3]. The speech recognition system is trained to recognize ten different words: the Portuguese correspondent words for the numbers "ZERO" to "NINE". The training set is composed of samples of each word, spoken by male speakers, with ages ranging from 20 to 55 years old.

This article is organized in 5 additional sections. Section 2 presents the speech recognition system, describing the signal acquisition method, the endpoint detection algorithm [4] and the time alignment technique [5].

Section 3 describes the speech analysis models, based on the extraction of spectral vectors from the speech samples. Some theoretical aspects, involving the different type of coefficients that form the spectral vectors, are presented.

Section 4 describes the Neural Network modeling for the speech recognition system. It describes the architecture and the learning algorithm employed.

Section 5 presents the tests performed, comparing the results obtained from the different spectral analysis models.

Section 6 presents a discussion on the results obtained and the conclusions of the work.

2 . SPEECH RECOGNITION SYSTEM

The speech recognition system is composed of the following modules:

1. ***Speech Signal Acquisition;***
2. ***Endpoint Detection and Time Alignment;***
3. ***Spectral Analysis Model;***
4. ***Artificial Neural Network (ANN).***

Figure 1 shows a schematic of the speech recognition system.

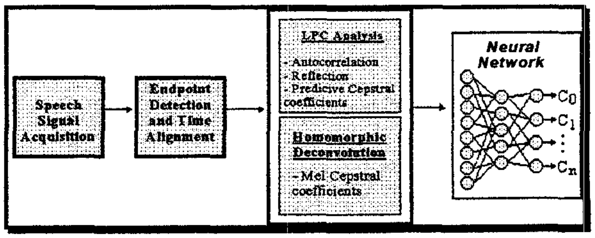
Speech Signal Acquisition
Endpoint Detection and Time Alignment
LPC Analysis
- Autocorrelation
- Reflection
- Predicive Cepstral coefficients
Homomorphic Deconvolution
- Mel Cepstral coefficients
Neural Network
C_0
C_1
:
C_n

Figure 1 - Schematic of the Speech Recognition System

The speech signal acquisition, the endpoint detection and the time alignment modules are described below. The two other modules, Spectral Analysis Model and Artificial Neural Networks, are discussed in the next sections.

Speech Signal Acquisition is performed through the A/D interface of a SUN SPARC workstation, with the speech signal being sampled at 8KHz. The data was recorded in a relatively noisy environment in order to test the robustness of the Neural Network system.

Endpoint detection is a method for eliminating the silence before and after each utterance, so that only relevant data is analyzed in the system's subsequent

0-7803-3636-4/97 $10.00 © 1997 IEEE

phases. An algorithm based on *the short time energy* and *the zero crossing rate* of the speech signal is used for this purpose [4].

This algorithm assumes that during the first 100 ms of the recorded signal, there is no speech present. Therefore, during this interval, the statistics of the background sound, which are the energy and the zero crossing rate, can be measured. Afterwards, a 10 ms window slides through the speech signal and these two statistics are computed in each step. Then, the beginning and the end points of the utterance are found by setting up thresholds for the values of the statistics measurements. Further details of this algorithm can be found in [4].

The *time alignment* problem derives from the fact that different speakers will say the same word in different speeds, due to different speaking rates [5]. However, the spectral sequences obtained to train and test the neural network have to satisfy certain order constraints. This means that the first ***n*** elements of every spectral sequence of an utterance of a word must correspond to the same phoneme. In order to handle this problem, we have used *windows of adaptive size* to perform a short time analysis. Each window defines a frame in the speech signal and each frame is the basic unit for the spectral analysis. By fixing both the number of windows that slide through the utterances and the superposition of adjacent windows, the correct duration of the sliding window can be calculated based on the total duration of the utterance. We have fixed a number of 150 *windows* and a *50%* superposition of adjacent frames. The average duration of each frame is about 20ms (corresponding to 160 speech samples), but as stated previously, it varies according to the duration of the utterance. Since the speech signal may be considered constant in the 10ms to 40ms interval, it may be assumed that no relevant information is lost with this approach. Figure 2 illustrates the time alignment technique used.

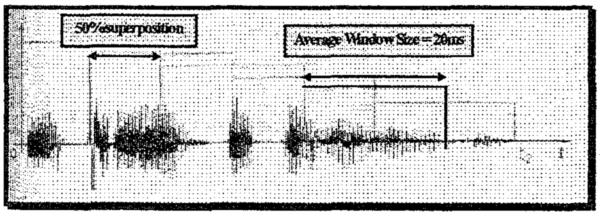

Figure 2 - Windows of Adaptive Size Used in the Signal Analysis

3 . SPEECH ANALYSIS MODELS

Spectral analysis models are used to compose the feature vectors of the utterances that will be submitted to the Neural Network. Two main analysis techniques have been used in this work: *Homomorphic Deconvolution [6]* and *Linear Predictive Code (LPC)[2,7]* analysis.

In order to use Homomorphic Deconvolution, the speech signal *s(n)* is represented in the time domain by the convolution of the excitation signal *u(n)* and the vocal tract response *h(n) [7]*, as shown in Equation (1).

$$s(n) = u(n) * h(n) \qquad (1)$$

The goal of the homomorphic deconvolution is to separate the signals *u(n)* and *h(n)* by transforming the speech signal *s(n)* into the so called *cepstron domain*. After the separation, the vocal tract response signal, in the cepstron domain, is used to form the feature vectors. The use of samples of the vocal tract response signal *h(n)* to form the feature vectors provides the desired speaker invariance. Each value of these vectors is called a Cepstron or MelCepstron coefficient.

The Linear Predictive Code is another approach to form the feature or spectral vectors. The idea behind the LPC model is that a given speech signal *s(n)* can be approximated by a linear combination of the past samples [2], as shown in Equation (2).

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots + a_p s(n-p) \quad (2)$$

where each coefficient a_i is obtained through the autocorrelation of the speech signal. An advantage of the LPC analysis is that, through appropriate conversions, other sets of coefficients, such as the *Reflection, the log Area Ratio* and *Linear Predictive Cepstral* coefficients [2], may be generated. These new sets of coefficients are called the *LPC parameter sets*. The formal method for converting autocorrelation coefficients into a LPC parameter set is known as *Durbin's method [2]* .

We have used the OGI Speech Tools [8], developed by the University of Oregon, to compute the three different types of coefficients. Once the coefficients are extracted, they are presented to the neural network to execute the training process.

4 . NEURAL NETWORK MODELING

Artificial Neural Networks [9] have emerged as a promising approach to the problem of Speech Recognition [10]. Using ANNs to handle Speech Recognition has the following advantages [2]:

- ***they are able to execute a high level of parallel computation;***
- ***they have a high level of robustness and fault tolerance capability;***
- ***they can learn complex features from the data, due to the non-linear structure of the artificial neuron [9, 11].***

The *Neural Network training algorithm* is another important component of the speech recognition system. In this work, the *Back Propagation [1,9]* algorithm has been selected to train the neural network. Although other Neural Network models, such as Time Delay Neural Networks (TDNN) [12], have been employed in Speech Recognition, this work shows that the traditional Back Propagation algorithm performs well for word recognition [10]. The combination of adequate endpoint detection and time alignment strategies provides the means to handle the speech dynamics without increasing the Neural Network complexity. The main advantage of this approach is the simplicity of the processing element (artificial neuron) when compared to other Neural Network approaches.

The Neural Network architecture is mainly determined by the number of frames in each utterance and by the size of the spectral vector representing each frame. In this work, the utterances are divided into 150 frames and the dimension of the each spectral vector is about 5 (depending on the type of analysis). Hence, there are about 600 to 900 neurons in the input layer of the neural net. We have used one hidden layer with a number of neurons ranging from 50 to 200. There are 10 neurons in the output layer, each one representing a decimal digit.

5. COMPARISON RESULTS

In this section we present the results obtained using the following coefficients as the network inputs:

- *Mel Cepstron Coefficients;*
- *Auto Correlation Coefficients;*
- *Mel Cepstron and Auto Correlation Coefficients (Hybrid Approach);*
- *Reflection Coefficients.*

In order to evaluate the performance of the speech recognition system in each case, we have carried out two different tests:

1. Training the Neural Network with three samples of each word from 9 speakers and testing it with different speakers not present in the training set.
2. Training the Neural Network with two samples of each word from 9 speakers and testing it with the third sample of the same speakers.

Since we have 10 different words, there are 270 utterances in the training set for the first case and 180 utterances for the second case. In both cases, the ANN is tested with utterances that were not presented in the training set.

The following sections present the results for each case.

5.1 - Mel Cepstron Coefficients

Tables 1 and 2 show the respective results for the two tests defined above. Each table indicates the statistics obtained for five speakers.

Speaker	Recognition Rate
A	90%
B	70%
C	60%
D	70%
E	70%

Table 1- Test 1 Results for Mel Cepstron Coefficients

Speaker	Recognition Rate
F	90%
G	70%
H	100%
I	70%
J	80%

Table 2 - Test 2 Results for Mel Cepstron Coefficients

The ANN architecture was composed of 900 neurons in the input layer, 90 neurons in the hidden layer and 10 neurons in the output layer.

5.2 - Auto Correlation Coefficients

Tables 3 and 4 present the results using the autocorrelation coefficients.

Speaker	Recognition Rate
A	80%
B	70%
C	60%
D	80%
E	70%

Table 3-Test 1 Results for Auto Correlation Coefficients

Speaker	Recognition Rate
F	80%
G	70%
H	80%
I	80%
J	90%

Table 4-Test 2 Results for Auto Correlation Coefficients

In this case, we have used an architecture comprising 600 neurons in the input layer, 100 neurons in the hidden layer and 10 neurons in the output layer.

5.3 - Hybrid Model

In this case, the ANN has 750 inputs - 450 are Mel Cepstron coefficients and 300 are Auto Correlation coefficients, and the hidden layer contains 80 neurons. Tables 5 and 6 show the results.

Speaker	Recognition Rate
A	80%
B	70%
C	80%
D	60%
E	80%

Table 5- Test 1 Results for the Hybrid Model

Speaker	Recognition Rate
F	80%
G	70%
H	90%
I	90%
J	90%

Table 6 - Test 2 Results for the Hybrid Model

5.4 - Reflections Coefficients

Tables 7 and 8 present the results when using reflection coefficients.

Speaker	Recognition Rate
A	80%
B	60%
C	80%
D	70%
E	70%

Table 7- Test 1 Results for Reflection Coefficients

Speaker	Recognition Rate
F	90%
G	70%
H	100%
I	70%
J	70%

Table 8 - Test 2 Results for Reflection Coefficients

We have used an ANN with 600 neurons in the input layer, 100 neurons in the hidden layer and 10 neurons in the output layer.

5.5 - Comparison

Table 9 summarizes the results given in the previous sections.

Approach	Test 1 Average Results	Test 2 Average Results
Mel Cepstron	72%	82%
AutoCorrelation	72%	80%
Hybrid	**74%**	**84%**
Reflection	72%	80%

Table 9 - Average Results Summary

Table 9 shows that the four approaches have very similar performance. However, it can be seen that the hybrid approach presents slightly better results, indicating the advantage of combining different sets of coefficients. In this particular case, the two sets of coefficients combines features derived from the homomorphic deconvolution and from the autocorrelation analysis.

6. CONCLUSIONS AND FINAL REMARKS

The Backpropagation ANN model has been applied to the problem of recognizing isolated words. The main objective has been to compare the performance of the recognition system with different sets of coefficients. The results obtained when using Auto Correlation, Mel Cepstron and Reflection coefficients were very similar, with an average recognition rate of 82% in the test 2 (train and test with the same speakers). The results obtained when using two different coefficients, the Mel Cepstron and the Auto Correlation coefficients, indicate that the hybrid model is the best approach. However, a more detailed investigation should be performed in order to come to a final conclusion.

The ANN performance can be considered satisfactory if we consider that:

- The speech acquisition environment has no noise control facilities (like in a real application);
- The speakers have a very different range of ages.

The training time in a SUN SPARC workstation was about 24 hours for all cases. The high number of neurons in the input layer contribute to an increase in the training time.

Some tests have also been performed, using *two hidden layers*. In this case, the training time drastically increased with no improvement in the system performance.

Acknowledges

We wish to thanks the Center for Spoken Language Understanding of the Oregon University for the speech tools software and the engineers Sidney C. B. Dos Santos and Roberto Garcia, PUC-Rio, Brazil, for their collaboration.

References

[1]-Jacek M. Zurada, "Introduction to Artificial Neural Systems", West Publishing Company, 1992.

[2] - Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice-Hall, 1993.

[3] - Yau-Hwang Kuo, Cheng-I Kao, Jiahn-Jung Chen, "A Fuzzy Neural Network Model and its Hardware Implementation", IEEE Transactions on Fuzzy Systems, Vol. 1, No. 3, August, 1993.

[4] - L. R. Rabiner and M. R, Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", The Bell system Technical Journal, Vol. 54, No. 2, February 1975.

[5] - Marconi dos Reis Bezerra, Pedro Paulo Leite do Prado and Sidney C. B. dos Santos, "Reconhecimento Automático de Locutor Utilizando Técnicas de Redes Neurais", Revista de Ciência e Tecnologia do Exército (Technology and Science Journal of the Brazilian Army), Vol. XII, No. 2, 2nd. Trimester, 1995.

[6] - A. V. Oppenheim and R. W. Schafer, "Discrete-Time Signal Processing", Prentice-Hall Signal Processing Series, 1989.

[7] - L. R. Rabiner and R. W. Schafer , "Digital Signal Processing of Speech Signals", Prentice-Hall Signal Processing Series, 1978.

[8] - Speech Tools User Manual, Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, 1993.

[9] - Philip P. Wassermann, "Neural Computing: Theory and Practice", VNR, New York, 1989.

[10] - Richard P. Lipmann, "Review of Neural Networks for Speech Recognition", Neural Computation 1, pp.1-38, Massachusets Institute of Technology, 1989.

[11] - R.S. Zebulum, K. Guedes, M. Vellasco, M.A. Pacheco, "Short Term Load Forecasting Using Neural Nets", Proceedings of the International Workshop on Artificial Neural Networks, LNCS No. 930, Springer-Verlag, Torremolinos, Spain, June,1995.

[12] - A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks" , IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No. 3, pp. 328-339, March, 1989.